User Variance and its Impact on Video Retrieval Benchmarking

Peter Wilkins[1], Raphaël Troncy[2], Martin Halvey[3], Daragh Byrne[1], Alia Amin[4], P. Punitha[3], Alan F. Smeaton[1], Robert Villa[3]
[1]CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland
[2]EURECOM, 2229, route des Cretes, Sophia Antipolis, France
[3]Computing Science Dept., University of Glasgow, UK
[4]CWI Amsterdam, The Netherlands
pwilkins@computing.dcu.ie

ABSTRACT

In this paper, we describe one of the largest multi-site interactive video retrieval experiments conducted in a laboratory setting. Interactive video retrieval performance is difficult to cross-compare as variables exist across users, interfaces and the underlying retrieval engine. Conducted within the framework of TRECVID 2008, we completed a multi-site, multi-interface experiment. Three institutes participated involving 36 users, 12 each from Dublin City University (DCU, Ireland), University of Glasgow (GU, Scotland) and Centrum Wiskunde & Informatica (CWI, the Netherlands). Three user interfaces were developed which all used the same search service. Using a latin squares arrangement, each user completed 12 topics, leading to 6 TRECVID runs per site, 18 in total. This allowed us to isolate the factors of users and interfaces from retrieval performance. In this paper we present an analysis of both the quantitative and qualitative data generated from this experiment, demonstrating that for interactive video retrieval with "novice" users, performance can vary by up to 300% for the same system using different sets of users, whilst differences in performance of interface variants was in comparison not statistically different. Our results have implications for the manner in which interactive video retrieval experiments using non-expert users are evaluated. The primary focus of this paper is in highlighting that non-expert users generate very large performance fluctuations, which may either mask or create system variability. The discussion of why this happened is not covered by this paper.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Evaluation/methodology, Video*; H.3.3 [**Information Storage and Retrieval**]: Information search & retrieval—*Information filtering, search process*

CIVR '09 July 8-10, 2009 Santorini, GR

General Terms

Experimentation, Human Factors, Measurement

Keywords

TRECVID, CBMIR, Video Retrieval, User Study

1. INTRODUCTION

Interactive video retrieval performance is difficult to assess due to a variety of factors, such as the effect of the 'user in the loop', search expertise or aptitude of the user, the graphical interface and the retrieval engine. The interplay between these various factors has always been difficult to disambiguate, particularly within benchmarked video retrieval evaluations which favor reporting of mean average precision - a measure of system performance - rather than human performance measures. In TRECVID [16], attempting to disambiguate these factors which may affect the performance of interactive retrieval is difficult, as only a list of saved 'shots' is returned by systems. Motivated by this and as part of the TRECVID 2008 interactive video search task, the K-Space[1][2] group undertook a novel experiment: conducting a cross-site evaluation using three different search interfaces with a common search engine. In total 36 users, 12 from three different sites, were employed to perform searches using each interface. By including multiple geographic sites it diversified our user base, whilst decoupling the retrieval engine from the interfaces allowed each interface to have standard and consistent performance. This facilitated the examination of both the user and interface effect and the extent to which these factors may impact on retrieval performance independent of algorithmic performance.

To the best of our knowledge, a content-based interactive video retrieval experiment within laboratory conditions of this size has not previously been undertaken. This experiment demonstrates that by providing a common search engine to multiple user interfaces, whilst gathering quantitative and qualitative metrics from participants, significant insights can be obtained into the factors influencing retrieval performance. We find through the use of transparent cross-site human performance measurements, that the largest factor determining search performance is the users chosen to

[1]K-Space is a European Network of Excellence (NoE) in semantic inference for semi-automatic annotation and retrieval of multimedia content
[2]Raphaël Troncy participated in this work whilst at CWI.

perform the task. We note that the effect of a 'human in the loop' may be extremely unpredictable (Section 5) and as such, there is definite need to share such measures, as they are key to understanding the factors which determine the reported success of a given interactive search system.

This paper is setup as follows. We present related work in the following section. The retrieval engine and interfaces are outlined in Section 3. The background and expertise of the users was carefully documented as well as their perception of the interfaces and search sessions and is presented in Section 4. Section 5 presents our results and analysis of the outcomes of our experiment. We finish with a discussion of the results and their potential impact on benchmarking activities in Section 6.

2. RELATED WORK

Within the information retrieval (IR) community, a number of evaluation and benchmarking activities have been established. These share the common goal of providing a large scale collection of data in order to achieve 'benchmarked' or comparable evaluation across various sites. Benchmarking provides a standardized, metricated evaluation to enable the comparison between information retrieval systems based on performance. Several benchmarking efforts exist, including TREC and CLEF, but perhaps one of the best known multimedia initiatives is the TREC Video Retrieval Evaluation (TRECVID) initiative [16]. Since its inception in 2000, the National Institute of Standards and Technology (NIST) has coordinated this activity annually. The goal of this benchmarking activity is *'to promote progress in content-based retrieval from digital video via open, metrics-based evaluation'* [16]. In so doing, TRECVID encourages research on multimedia information retrieval by providing a large test collection, uniform scoring procedures, and a forum for the comparison of results. As part of the evaluation, participants are provided with a development and test corpus of broadcast video footage. Each team then builds a retrieval system using the development data to guide them and performs a series of topic-based experiments. In the interactive search task, a user is provided with a graphical user interface and must complete a set of search topics. For each topic, the user is provided with a set of visual exemplars and descriptive text e.g. "*Find shots of one or more people walking up stairs*" and the search must be completed within 10 minutes. In the 2008 task, users could perform 24 topics and teams were allowed to submit up to 6 runs. The human judged relevant items were then reported to NIST and validated. The interactive search task is particularly important for two major reasons. First, it aims to replicate a real world scenario in which the searcher can react to the search results by, for example, reformulating the queries. Second, it is evident that video retrieval with a 'human in the loop' will far outperform any automatic methods. It is as such essential to not only understand the affect of the underlying retrieval engine when discussing system performance but also to quantify the role of the user, the interface and their interplay [5].

Interest in interactive retrieval is widespread across the Information Science domain. Within the text retrieval community, there have been multiple efforts to attempt to disambiguate variables which contribute to retrieval performance. One of the most notable of these activities was the Interactive Track of TREC, beginning in TREC-6 [11]. The objective of this activity is the same as ours, "*isolating the effects of topics, human searchers, and other site-specific factors*" [11]. Similar to what we have attempted, this activity required that participating sites not only record documents that user's saved, but also to record complex interaction logs, demographic data and qualitative metrics. However, the key advantage of this activity which we currently do not have in TRECVID, was that participants in the Interactive Track were required to run in conjunction with their own retrieval systems a baseline retrieval system supplied by NIST. This allowed for a comparative evaluation of the abilities of the searchers involved at each site, thus allowing comparisons of systems which took this into consideration. This motivation directly applies to our work, as we seek to replicate this scenario by having all three user interface variations being used at each site with each interface using a common search engine, giving us an idea of the variance within our user set. The objective of having a baseline retrieval system has always been an intention of NIST [15]. However given the complexities of content-based multimedia information retrieval systems this has proven difficult to achieve.

To the best of our knowledge, no efforts have been made to share interaction data from TRECVID or other multimedia initiatives. However, the Open Video Digital Library (OVDL) has provided a repository of digital content, and an open interface for browsing and searching the data [12]. Despite the lack of shared interaction data from TRECVID evaluations, the common collections and benchmarking resources provided by this initiative have facilitated a great deal of research into both interactive retrieval and its associated human factors. For example, MediaMill at the University of Amsterdam build rich visualizations of the result-space (e.g. the Fork-, Cross- and Rotor-Browser) that enable users to easily explore the full depth of often-complex result-sets [6]. The team from the National University of Singapore pushes the boundaries of 'extreme retrieval' by forcing the user to make judgments on a result's relevance within a very limited time window [13]. FXPAL has evaluated a collaborative retrieval system under the TRECVID benchmark [1].

Perhaps of most interest are the explorations conducted by researchers at Carnegie Mellon who have extensively explored user-centered issues in video retrieval. Christel and Conescu previously investigated how best to support the novice within the retrieval process through techniques such as shot suppression and by encouraging the use of different access mechanisms within a shot-based interface [5]. Interestingly, they show that the suppression of previously seen shots did not have the anticipated positive effect on performance. Christel has also discusses the distinctions between the novice and experts and outlines the design considerations required to cater for these roles within the retrieval process [3]. Additionally, Christel has considered the use of storyboards, a grid layout of thumbnail images as surrogates representing video for video search, a commonly-adopted metaphor within video retrieval interfaces. He remarks that such story boards offer many advantages in exploratory, shot-based retrieval but moving forward, support for longer term search activities needs to be considered [4]. Hauptmann and Christel [8] have also surveyed the 'state of the art' in TRECVID search systems discussing the features which contribute to the success within an interactive

retrieval. They highlight the importance of text retrieval noting it to be "*much more robust than any of the visual features*". Moreover, they highlight the utility of temporal context in interactive retrieval, a topic which Yang and Hauptmann have further explored as a means by which to augment the ranking of search results [8]. They define temporal consistency as "*the tendency that the relevant shots ... appear in temporal proximity*" for a given semantic concept or query. They note that while the degree to which relevant items are temporally proximal is dependent on the topic, temporal context is extremely useful in video retrieval.

Building upon some of this prior work, the K-Space group conducted an interactive retrieval experiment as part of TRECVID 2007 [2]. The investigation was designed to further explore the role of temporal context within interactive search. This was achieved by creating two search interfaces which offered the polar extremes of temporal context and by logging all user interactions throughout participants search sessions. The first variant was recall-oriented, offering a large number of results without any context information while the second was context-oriented by placing each result in the context of the full broadcast. Apart from sharing the same retrieval engine, both systems also shared a common query input panel, topic description panel and saved shot area. The only major difference was in the presentation of the results from the underlying retrieval engine. Furthermore, the affect of context-provision was explored for both novice and expert searchers. While performance in both systems was comparable, with experts notably outperforming novices, the progression of the search and the search strategies adopted by the users was markedly different for each interface. Interestingly, users of the recall-oriented system often failed to find relevant temporal siblings for a relevant shot. As such the authors suggest that the presentation of some temporal context within shot-based interfaces can be used to significantly and effectively augment the number of relevant items while minimizing user effort (where effort is search reformulation).

3. SYSTEM DESCRIPTION

The three user interfaces developed for the search experiment leveraged a common search engine that makes use of several content-analysis techniques. We briefly detail these, with a more complete explanation in [17].

As no common keyframe set was released by TRECVID, we extracted our own set of keyframes. Our keyframe selection strategy was to extract every second I-Frame from each shot. We extracted low-level visual features from K-frames using several feature descriptors based on the MPEG-7 XM. These descriptors were implemented as part of the aceToolbox[3], a set of low-level audio and visual analysis tools developed in the EU aceMedia project. We made use of six different global visual descriptors. These descriptors were Colour Layout, Colour Moments, Colour Structure, Homogeneous Texture, Edge Histogram and Scalable Colour.

The common search engine leveraged multiple modalities to form a response to an information need from a user. The search engine allows for multiple query by example, text queries and mixed modality queries. For visual components of queries we made use of six global visual features identified earlier, ranking within each was handled by the similarity measures as specified by the MPEG7 specification. These measures for the most part are similar to Euclidian distance. Our retrieval engine also made use of High-Level Features (i.e. concepts), which were generated by the K-Space partners and covered the 36 semantic features required for participation in TRECVID 2007. Further details can be found in [17].

The previous content-analysis techniques could be accessed via two mechanisms within the search engine. The first method was to use the outputs of the previous methods as 'filters' on a result set of shots. The filters could have three states, 'show only shots matching the filter', 'shots not matching the filter' and no effect (default).

The second method of access incorporated not only the K-Space content-analysis results, but also results from the CU-VIREO374 collection donated by City University of Hong Kong and Columbia University [10] for which we are very grateful. We took the names of the concepts detectors and ran these through Wordnet obtaining the synonyms for these terms. Therefore for each shot we had a bag of words which described the visual aspects of that shot. This text for each shot was then augmented with the translated ASR text provided by the University of Twente [9]. This therefore produced for each shot a collection of terms which described the content of the shot incorporate both visual and audio information. The text was then indexed by Terrier [14], with retrieval results provided through a vector space model.

3.1 Three interfaces for the Interactive Search

The following subsections provided an overview of the user interfaces. Further description can be found in [17].

3.1.1 Shot based Interface (DCU-1)

The 'shot based' system presented to the user the ranked list of shots direct from the retrieval engine. The ranked shots are organized left to right, top to bottom (Figure 1). It can be thought of as the more traditional result display that has been used for content-based retrieval interfaces. This interface displays no context for any of the returned results.

Figure 1: Shot-based user interface

3.1.2 Broadcast based Interface (DCU-2)

The 'broadcast based' system takes the idea of context to its extreme by ranking not shots, but broadcasts. The magazine/documentary broadcasts which compose the TRECVID 2008 corpus tend to be about one major subject, whilst in

[3] https://kspace.cdvp.dcu.ie/secure/aceToolbox.zip

previous years a news broadcast could be seen as containing many subjects. With this in mind we can assume that the shots within a broadcast are more homogeneous. As such ranking broadcasts as opposed to shots appears as an interesting alternative. In Figure 2, we can see a horizontal line of shots in rows across the results area. Each of these rows is a ranked entire broadcast, with the best-matching broadcast being the first row. When a user issues a query, the ranked list of broadcasts is presented, and within each broadcast's row the row will be centered on the highest matching shot within that broadcast.

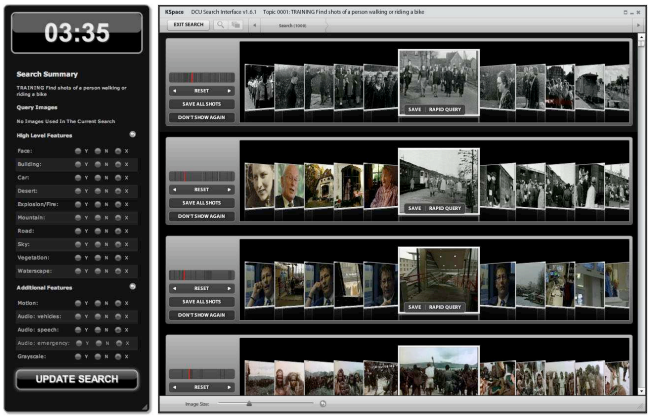

Figure 2: Broadcast-based user interface

3.1.3 *Zooming Interface (GU)*

The Zooming interface leveraged temporal context as well as a diversity re-ranking of the search results. The value and importance of a search result appears to be based on it's value as a good starting point for a user to find other relevant shots within a video, by browsing the video, as much as the relevance of the result itself. Based on the ability of users to easily browse videos, and the willingness of many users to do so in order to find relevant material, we constructed an interface that (a) emphasizes results provide are good starting points from which to find material (point-finding within videos), and (b) extend the video browsing elements of the user interface to enable users to more easily view and browse videos. To address (a) we introduced a diversity based re-ranking to the search results which offers more 'starting off points' for browsing. In order to achieve (b) a zooming interface was implemented to aid users in exploring more of a videos content when engaged in neighborhood search.

4. EVALUATION

We conducted our interactive video retrieval evaluation under laboratory settings, carrying out the experiment in three geographically different locations (CWI in Amsterdam (NL), DCU in Dublin (IE), GU in Glasgow (UK)). Each participant used the 3 different search interfaces described above: Shot-based, Broadcast-based, and Zooming interface (within subject design). In total, they were required to complete 12 video search tasks (4 video search topics per interface) taken from the TRECVID 2008. Users completed a training session prior to the main task to ensure they were familiar with the interfaces operations and functionality and that they fully understood the search tasks. A participant needed to complete a current task before proceeding to the next one.

Figure 3: Zooming user interface

The time given to complete a task was 10 minutes. Micro breaks were introduced between tasks to allow participants to refresh themselves. Participants were given a questionnaire during the experiment, with background information on the individual and their search experience collected prior to training. For each topics, users interactions were logged (time, video searched and browsed, video played, video saved and video removed), providing an extensive amount of information on the participants interaction with the system. Following each task, participants were asked to appraise their performance, while at the completion of tasks for an interface, they were required to assess the system. The questionnaire were based on the AttrakDiff questionnaire [7] and probed the hedonic and ergonomic aspects, usability, and positive and negative experiences of using the system. After the evaluation, we conducted debriefing interviews with the participants to gain more informal feedback.

All users were instructed to save as many shots as possible that matched the TRECVID topic description. If a user was unsure, it was left to his/her discretion whether to save the shot or not, but the emphasis of the task was to find as many matches as possible.

In total 36 people participated in our study. They are students or researchers, recruited equally from the 3 different institutions mentioned above. Participants are mostly male (75%), aged between 25 to 56 years old (M=29.1, SD=6.2). While most participants were experienced searchers, they are novice video searchers and occasionally use video searching applications (see Table 4). A few (4 from 36 people) were advanced video searchers and frequently use video searching applications. We anticipated that some population bias would be in effect, the participants from DCU were from our research group and would be more familiar with the concepts of content-based retrieval, than graduate students from CWI whose specialties lie elsewhere.

5. RESULTS

In this section, we will present an analysis of our user experiment, specifically examining issues concerning users and their impact in retrieval performance, from both quantitative and qualitative perspectives. This analysis will demonstrate that given the same retrieval engine and user inter-

Age:	22-56 years old, M=29.1, SD=6.2
Gender:	Male (27), Female (9)
Education:	graduate students (13), researchers (17), other (6)
General search exp.*:	M=4.6, SD=0.9
Video search exp.*:	M=2.0, SD=1.1
Affiliation:	DCU (12), CWI (12), GU (12)

* 1:none, 3:fairly (1 search daily), 5:very frequently (several daily)

Table 1: Participants demographic (Total: 36)

System	InfAP	P10
DCU-1 (Max)	0.0366	0.4125
DCU-2 (Max)	0.0306	0.3750
GU (Max)	0.0366	0.3750
DCU-1 (Med)	0.0123	0.2104
DCU-2 (Med)	0.0077	0.1437
GU (Med)	0.0121	0.1875

Table 2: 2008 Interactive Search Results

faces, the performance of individual users has far greater impact on search performance than previously anticipated.

Table 2 presents an analysis of our results using the TRECVID evaluation metrics of Inferred Average Precision (InfAP) and Precision@10 (P10). For each system we present two runs: 'max' and 'med'. We evaluated every user's performance for every topic to compose these runs. The 'max' run is the selection of topic results which achieved the best performance in terms of InfAP for that system. The 'med' run is the median run, where the selection of results was obtained by calculating the median InfAP value for every topic for each system.

The results presented in Table 2 startling illustrate the impact of user selection in comprising runs for submission to a benchmarking activity, and the degree to which user's performance varies. Each of our 'max' runs obtains three times the performance of it's equivalent median run. This result was unexpected, whilst we anticipated some variability in our results due to differences in user populations, the magnitude of the observed difference is alarming.

Evaluating these runs, we ran significance tests using a $\rho = 0.01$ and found that whilst every 'max' run was significantly better than the 'med' runs, that within each 'class' of run (i.e. 'max' and 'med') there was no significant difference (i.e. all 'max' runs were not significantly different to each other, likewise for the 'med' runs). On the one hand, this means that given equivalent sets of users, each of the interfaces performed at approximately the same level. However within the same system we note the massive discrepancy between the performance of the 'max' and 'med' runs. From the same set of users, we were able to produce representative runs which varied wildly, indicating large performance variance in our user set.

To investigate this further, we examined which shots a user saved for each topic for each system, to determine if there was any commonalities. We utilize the Jaccard index which provides a measure of how similar or dissimilar two sets are. A Jaccard index value is in the range [0:1], a value of 0 means that the sets are mutually exclusive, a value of 1 means that the sets are the same. We compute for each topic the Jaccard index within each system (DCU-1, DCU-2, GU) across sites, and compare this against the Jaccard index within each site (CWI, DCU, GU). The first comparison gives us an indication if users using the same system save similar shots, whilst the second tells us that within a location (e.g. CWI) if users for a topic are saving the same shots regardless of system used. The results are plotted in Figure 4.

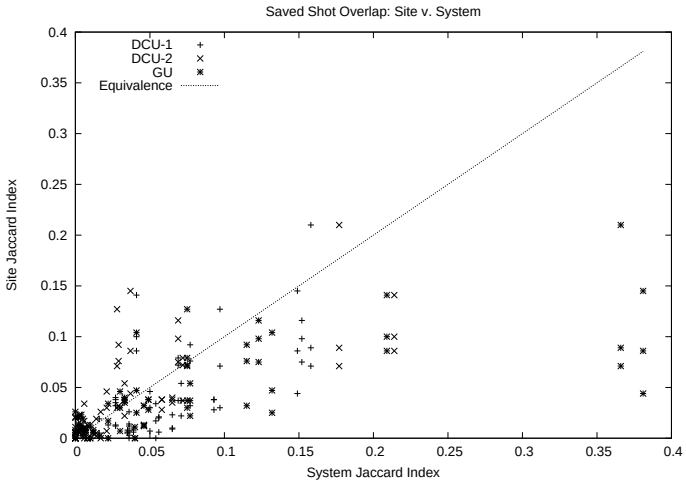

Figure 4: Saved Shot Set Overlaps

Points which occur on the line in the graph have equal values for system and site set overlaps. Points which occur to the right of this line have greater overlaps due to the system being used, whilst points to the left have greater overlaps due to the site. The points are labeled according to the system used. Our first observation is that for the majority of topics, there is very little overlap in the shots saved by users, as the majority of points lie in the range [0.05: 0.05], meaning that there is great variability in the shots selected as relevant by our users. However we do see some artifacts in the graph, for example the GU system (the star points) has multiple points to the right of the line, indicating that users of the GU system for certain topics were more likely to save the same shots, regardless of site. Alternatively, we see that the DCU-2 system for certain topics features points on the left of the line, indicating for those topics that users within a site using that system found the similar shots. As this interface promoted browsing the collection, groups (such as DCU) who have previous experience with content-based retrieval may have examined the collection in similar methods, resulting in a higher site overlap. The purpose of this graph was to establish if there was any commonality in the sets of shots the users saved, and if a bias could account for this (i.e. did users using the same system save the same shots, did users from the same site save the same shots). However on the whole there was little intersection of the shots users saved, lending further evidence towards the indications of massive variability of performance in our user base.

An alternative method for examining the variability of user performance is to examine the amount of shots saved by each user. In examining this, we make the assumption that if a user saved a shot, that for that user the shot is relevant. For every topic we determine the average number of shots saved by each system. Transforming the number of

shots saved for each topic by each user into a Z-Score, we are able to express for every topic how close or far the number of shots saved by a user was in comparison to the mean. We aggregated this data together at the site level. This allows us to express for a given site, how its users performed on average with regards to the average performance overall. Figure 5 displays these graphs, one for every site. On the X-axis of this graph are listed the standard deviations, $+2\sigma$ indicates that users were saving twice as many more shots than the average, while -2σ is the opposite. Each bar on the graph represents the average amount of saves for a given system.

The data displayed here confirms our earlier observations about variance in our user population. We can see for the CWI site demonstrated in the first graph, that the data follows a normal distribution. The majority of the users for CWI are saving about the average number of shots per topic for each system (i.e. the bulk of the mass is located $+/-1\sigma$). However, this contrasts to both the DCU and GU sites where both exhibit a skewed distribution. In the case of DCU, the distribution is skewed to the left, indicating that on average compared to the other sites, users at DCU were saving more than the average number of shots for any given topic and system. Conversely the GU data is skewed to the right, showing that on average the GU users saved less than the average number of shots.

When the previous evidence is taken together, it presents an unexpected picture. We have found the variability of users performance to be far greater than expected. User's saved few shots in common despite using the same information systems for the same task. Likewise when we constructed entire 'retrieval' runs for evaluation, by taking the best performing result for a topic and the median result, that the difference between these runs was approximately 300%, yet in each case runs of the same class were not statistically different. We also gathered qualitative data to obtain further insights into the search session.

5.1 Qualitative Data

With a a large number of users and consequently a large number of the same topics completed by users, there is an obvious challenge in both constructing retrieval runs for evaluation. One approach is to conduct the saved shot analysis as outlined above, while an alternative is to survey the users to gain a subjective measure of performance with an interface for any given topic. As part of the evaluation our participants were asked to gauge this after the completion of each topic. This offers us the ability to compare actual performance with perceived performance and to validate its potential utility in the composition of retrieval runs. The results of this are presented in Figure 6.

This graph aggregates each users estimate of their perceived performance per site for each of the interfaces (listed on the X-axis). The range for this data was [+3, -3]. From this, we can determine that all sites thought that they performed poorly using the DCU-2 system, which does correlate with the actual performance shown in Table 2. However the users from CWI and DCU expressed no perception of strongly positive performance for any of the systems, whilst users from GU thought that they performed well using the system developed within GU. This is indicative of a potential site bias present in the qualitative data and would suggest caution should be used in applying this data in selection

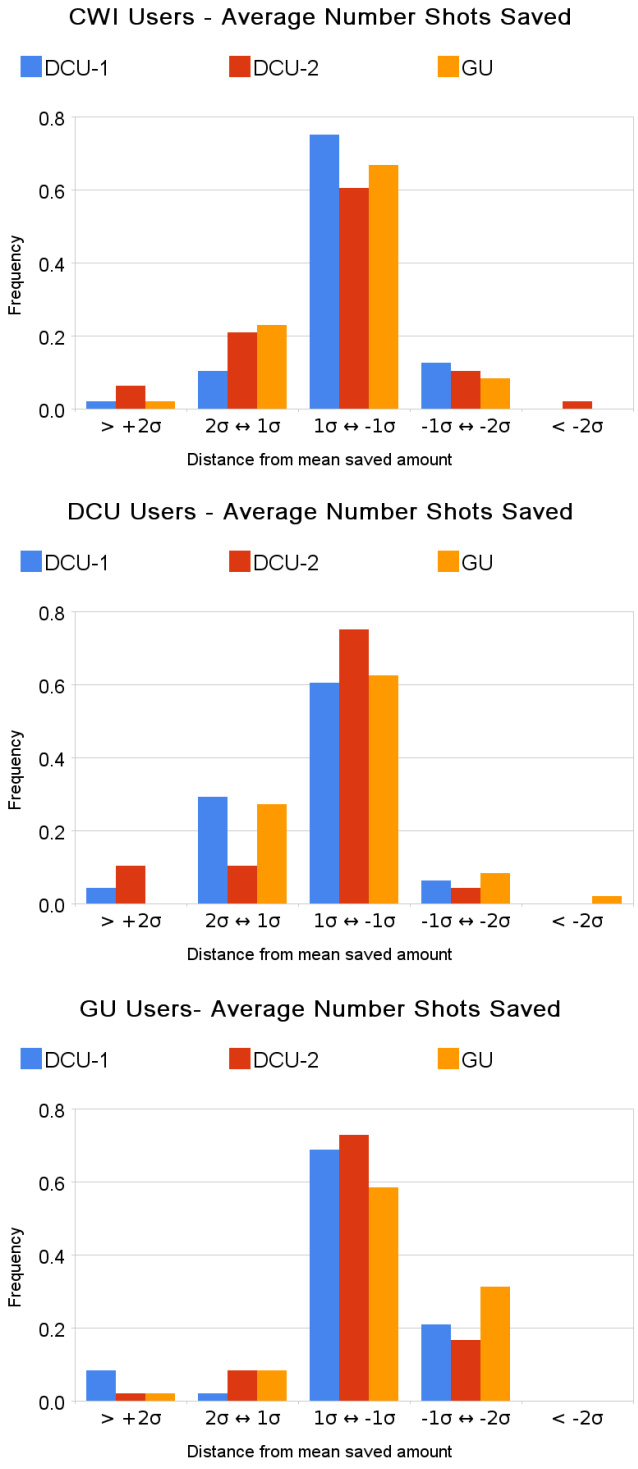

Figure 5: Average Number Shots Saved by Site

Table 3: Interface Comparison[1] (Total: 36 people)

	Mean Score (SD)			p.
Interface assessment:	Shot (DCU-1)	Broadcast (DCU-2)	Zooming (GU)	
Easy to use	0.87(1.36)	**-0.74(1.54)**	1.06(0.49)	F(2,33)=14.51, p<.05
Easy to learn	0.22(1.42)	-0.29(1.26)	**0.79(0.79)**	F(2,33)=9.20, p<.05
Ergonomic quality[2]	-0.72(0.66)	**0.07(0.99)**	-0.83(0.88)	F(2,33)=25.8, p<.05
Hedonic quality[2]	-0.58(-0.57)	-0.64(0.68)	**-0.03(-0.07)**	F(2,33)=6.84, p<.05
Appeal[2]	-0.74(0.72)	-0.25(1.01)	-0.35(1.10)	F(2,33)=3.01, p=.09
Self assessment:				
Overall performance using the interface	0.37(1.28)	**-1.50(1.17)**	1.42(0.97)	F(2,33)=36.43, p<.05

[1] Min:-3, Max:3

[2] AttrakDiff questionnaire [7]

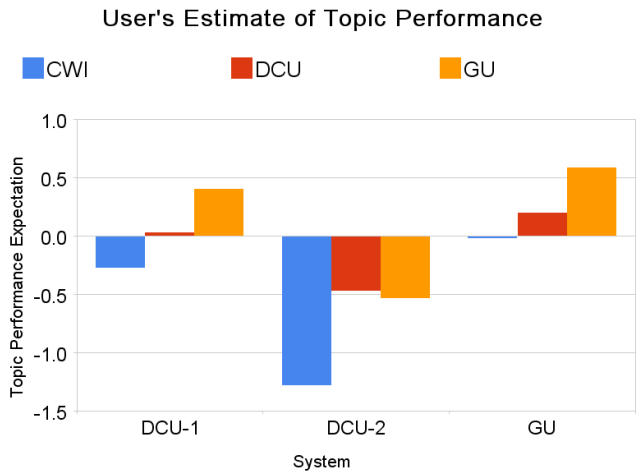

Figure 6: User estimation of topic performance for each system

strategies.

As mentioned in the Section 3, each of the three interfaces presented results in uniquely different format. In the post-system questionnaires, we solicited subjects' opinions on these interfaces and the techniques used to browse and present results. Participants were asked to rate their overall perceived performance with the interface, its ease of use, learnability, and its general appeal using 7-point Semantic differential scales, which yielded results in the range of -3 to +3. To assess the general appeal of the interface, users were administered twenty-three separate questions on 7 point semantic differential scale. These were based on the AttrakDiff questionnaire [7] and allowed the assessment of the user interfaces in three broad categories, namely: ergonomic quality; hedonic quality; and appeal. We applied two-way analysis of variance (ANOVA) to each differential across all 3 systems and the 24 topics to test the significance of these results, which are presented in Table 3. The mean value is displayed along with the standard deviation in brackets, values in bold are statistically significant.

From the results in Table 3, it appears that participants have a mixed reaction to the interfaces presented, with contrasting views particularly for the DCU-2 and GU systems. User's found that the DCU-2 system was the hardest to use, whilst the GU system was the easiest. The DCU-2 system however was deemed the most ergonomic of the interfaces, a potential artifact of the large amount of temporal context displayed. The GU system scored the highest for hedonic quality. Reinforced in the previous section, users found that the DCU-2 system was likely to result in poor performance.

6. DISCUSSION AND CONCLUSION

It is widely accepted that the performance of interactive video retrieval is impacted by many factors such as the interface, the retrieval experts utilized, the selection of keyframe extraction strategies and the user's employed to undertake the experiment. In this paper, we presented a large-scale retrieval experiment, making use of non-expert users, spread over three geographic sites. We found exceptionally highly levels of user variance in this activity. The fundamental implications of this reflect upon how we evaluate retrieval systems and if the conclusions we draw from experiments are robust.

We are certainly not the first to highlight this issue before, indeed many researchers have commented on this, and NIST itself would like these issues addressed as highlighted in TRECVID 2003 [4, 11, 15]. The conclusions of this paper may appear alarmist, calling into question observations reached from previous retrieval experiments in TRECVID as there was no user normalization. However, groups in recent years participating in TRECVID have avoided these complications by engaging in "expert" runs. That is the use of a single user who was involved in the creation of the retrieval system, but isolated from the test collection. This user is able to maximize the performance of the system they developed. We can then propose that groups who perform this style of interactive experiment are able to make more robust observations as they can compare against other "expert" systems.

The question of why our experiments elicited this large variability in user performance is an important question, one possibility raised by peer-review was the impact of topic complexity and its relationship to our observations. We conducted an examination of the correlation between our average user variance per topic and the overall TRECVID median average precision score per topic, finding a pearson correlation of 0.49, with variance greater when the median AP was high. Given that TRECVID 2008 was a low performing benchmark with regards to AP, this may account for some of our observations. However why we observed variance is not the focus of this work, the fact remains that large user

variance did in fact occur which leads us to question of how to interpret the resulting workshop outcomes.

Interactive video retrieval, both from an implementation and execution level (building a system and running an experiment) is undoubtedly hard. Within TRECVID, we are observing the rapid increase in popularity in fully automatic search, which with the user removed allows for robust comparisons of retrieval algorithms. Yet as video retrieval increases in popularity, it becomes ever more paramount for us to develop a better understanding of the human involved in the video retrieval context. The question becomes, what mechanisms can we employ to conduct non-expert retrieval experiments and achieve robust results?

The ideal solution would be what was employed in the TREC Interactive Track [11], a mandated baseline system which could be used to benchmark the users involved at each participants site. This though would require significant effort and is unlikely to happen in the near future. Other possibilities include the reporting of additional evaluation metrics [12], the move away from analyst-oriented deep information seeking tasks, or the change to more generalized corpora which are less dependent on "shots".

The activities of benchmarking evaluations such as TRECVID respond to the needs expressed by the community. We need to discuss what important factors should be being captured so that greater understandings of interactive video retrieval can be made. A simple beginning point would be the timestamping *or* inclusion in submitted results of only shots explicitly saved by a user in a search session. This would allow us to cross-compare even at a simple level how users varied across systems, which currently is not possible with many existing search results.

Acknowledgments

This paper was supported by the European Commission under contract FP6-027026 (K-Space) and by Science Foundation Ireland under grant 07/CE/I1147 (CLARITY: Centre for Sensor Web Technologies).

7. REFERENCES

[1] J. Adcock, J. Pickens, M. Cooper, F. Chen, and P. Qvarfordt. Fxpal interactive search experiments for trecvid 2007. In *TRECVid 2007 - Text REtrieval Conference TRECVid Workshop*, 2007.

[2] D. Byrne, P. Wilkins, G. Jones, A. F. Smeaton, and N. O'Connor. Measuring the impact of temporal context on video retrieval. In *CIVR 2008 - ACM International Conference on Image and Video Retrieval*, 2008.

[3] M. G. Christel. Establishing the utility of non-text search for news video retrieval with real world users. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 707–716, New York, NY, USA, 2007. ACM.

[4] M. G. Christel. Supporting video library exploratory search: when storyboards are not enough. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 447–456, New York, NY, USA, 2008. ACM.

[5] M. G. Christel and R. M. Conescu. Mining novice user activity with trecvid interactive retrieval tasks. In *CIVR*, pages 21–30, 2006.

[6] O. de Rooij, C. G. Snoek, and M. Worring. Balancing thread based navigation for targeted video search. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 485–494, New York, NY, USA, 2008. ACM.

[7] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner. Hedonic and ergonomic quality aspects determine a software's appeal. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 201–208, New York, NY, USA, 2000. ACM.

[8] A. G. Hauptmann and M. G. Christel. Successful approaches in the trec video retrieval evaluations. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 668–675, New York, NY, USA, 2004. ACM.

[9] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the second international conference on Semantics And digital Media Technologies (SAMT)*, Lecture Notes in Computer Science, Berlin, December 2007. Springer Verlag.

[10] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. Technical report, Columbia University, August 2008.

[11] E. Lagergren and P. Over. Comparing interactive information retrieval systems across sites: the trec-6 interactive track matrix experiment. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 164–172, New York, NY, USA, 1998. ACM.

[12] G. Marchionini, B. M. Wildemuth, and G. Geisler. The open video digital library: A möbius strip of research and practice. *Journal of the American Society for Information Science and Technology*, 57(12):1629–1643, 2006.

[13] S.-Y. Neo, H. Luan, Y. Zheng, H.-K. Goh, and T.-S. Chua. Visiongo: bridging users and multimedia video retrieval. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 559–560, New York, NY, USA, 2008. ACM.

[14] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in terrier. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007.

[15] A. F. Smeaton, W. Kraaij, , and P. Over. Trecvid 2003 - an overview. In *TRECVID 2003 - Text REtrieval Conference TRECVID Workshop*, MD, USA, 2003. National Institute of Standards and Technology.

[16] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

[17] P. Wilkins and et al. KSpace at TRECVid 2008. In *TRECVid 2008 – Text REtrieval Conference, TRECVID Workshop, Gaithersburg, Md., 17-18 November 2006*, 2008.